2. Currently Amended Claims and Status

We have amended the independent claims 1, 10, and 16 to add the material hardware elements to claim structure. Claim 7 and 12 are currently amended to improve clarity and narrowed with further limitations. Claims 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15 are currently amended for consistency with the independent system claim. Please let us know if this meets with your approval.

1. (currently amended) A system ~~and method of~~ for ~~communicating emotive content~~ processing emotive vectors comprising;

   at least one computing device,
   computer memory, and
   computing device communication medium

   whereby software instructions stored in memory are under control of the computing device for processing and transmitting emovectors over the communication medium, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author, with associated text embedded in electronic device communications.

2. (currently amended) A system ~~method~~ as in claim 1 further comprising the encoding of ~~emotive content~~ emotive vectors into standard computing device communication formats.

3. (currently amended) A system ~~method~~ as in claim 1 further comprising the encoding of the emotive content into textual communications.

4. (currently amended) A system ~~method~~ as in claim 1 further comprising the decoding of emotive content in electronic communications bearing emotive vectors normalized to the communication's author.

5. (currently amended) A system ~~method~~ as in claim 4 further comprising parsing the emotive content into tokens for presentation and display of face glyph emotive representations with associated textual content on receiver computing device displays.

6. (currently amended) A system ~~method~~ as in claim 5 further comprising the tokenizing of the parts of speech of associated text and with the tokenized emotive content synthesizing author's intended meaning text strings.

7. (currently amended) A system ~~method~~ as in claim 4 further comprising the mapping of emotive intensity numerical value ~~into~~ from one or more words, ~~text~~ from a pre-defined table of numerical values mapped to words ~~describing the emotive intensity value in express language which would qualify an associated emotive state with the intensity value~~.

8. (currently amended) A system ~~method~~ as in claim 1 further comprising the scanning and tokenizing of the embedded emotive content in the communications.

9. (currently amended) A system ~~method~~ as in claim 1 further comprising parsing communications containing the emotive content using emotive grammar productions to tokenize the emotive content in textual communications.

10. (currently amended) A method of encoding emotive vectors, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author with associated text in electronic communications, comprising the steps of:

reading the emotive vector into a computer memory from a computing device medium;

processing emotive vector at with least one computing device, and

transmitting the emotive vector to another computing device.

11. (original) The method in claim 10 further comprising structuring and synthesizing emotive parsers with productions exploiting emotive vectors encoded in textual datastreams.

12. (original) The method in claim 10 further comprising an emotive parser to tokenize emotive vectors into emotive components and emotive components to a set of face glyphs.

13. (currently amended) The method in claim 12 further comprising an emotive natural language parser to extract and tokenize emotive vector tokens decoupled from the associated natural language text ~~into the~~ parts of speech component tokens.

14. (original) The method in claim 13 further comprising concatenating communication tokenized emotive components with grammatical string fragments and strings selected from the associated text into grammatical strings conveying an intended meaning of the communication.

15. (original) The method in claim 14 further comprising said face glyph set based on graphic rendering of reasonably representative emotive states and associated emotive intensities.

16. (currently amended) A computer program residing on a computer-readable media, said computer program communicating emotive content comprising emotive vectors, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author with associated text embedded in electronic device communications, comprising the steps of:

reading the emotive vector into a computer memory from a computing device medium;

processing emotive vector with at least one computing device, and

transmitting the emotive vector to another computing device.

17. (currently allowed) A computer network comprising:

a plurality of computing devices connected by a network;

said computing devices which display graphical and textual output;

applications executing on the devices embedding emotive vectors which are representations of emotive states with associated author normalized emotive intensity;

assembling emotive content by associating emotive vectors with associated text in electronic communication;

encoding emotive content by preserving association of emotive vectors with associated text in the electronic communication;

transmitting the communication with emotive content to one or more receiver computing devices;

parsing communication bearing emotive content; and

mapping emotive vectors to face glyph representations from a set of face glyphs;

Such that communications encoded with emotive content facilitate exchange of precise emotive intelligence.

18. (currently allowed) A computer program residing on a computer-readable media, said computer program communicating over a computer network comprising:

a plurality of computing devices connected by a network;

said computing devices which display graphical and textual output;

computer-readable means for applications executing on the devices embedding emotive vectors which are representations of emotive states with associated author normalized emotive intensity;

computer-readable means for assembling emotive content by associating emotive vectors with associated text in electronic communication;

computer-readable means for encoding emotive content by preserving association of emotive vectors with associated text in the electronic communication;

computer-readable means for transmitting the communication with emotive content to one or more receiver computing devices;

computer-readable means for parsing communication bearing emotive content; and

computer-readable means for mapping emotive vectors to face glyph representations from a set of face glyphs; and

computer-readable means for displaying communication of textual with associated face glyph emotive representations on said computing device displays;

whereby communications encoded with emotive content provide means of exchange of precise emotive intelligence.

3. Previous Claim Status

Claims 1, 10, and 16 were amended reflect the definitions for emovector given in the specification on page 20, so they are expressly defined in the claims as per your request. Claim 17 was amended by striking 2 stray lines after the claim ending, making it not a part of the original claim 17 yet not part of claim 18. Claims 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, and 18 remained unchanged.

1. (previously amended) A system and method of communicating emotive content comprising emotive vectors, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author, with associated text embedded in electronic device communications.

2. (original) A method as in claim 1 further comprising the encoding of emotive content into standard computing device communication formats.

3. (original) A method as in claim 1 further comprising the encoding of the emotive content into textual communications.

4. (original) A method as in claim 1 further comprising the decoding of emotive content in electronic communications bearing emotive vectors normalized to the communication's author.

5. (original) A method as in claim 4 further comprising parsing the emotive content into tokens for presentation and display of face glyph emotive representations with associated textual content on receiver computing device displays.

6. (original) A method as in claim 5 further comprising the tokenizing of the parts of speech of associated text and with the tokenized emotive content synthesizing author's intended meaning text strings.

7. (original) A method as in claim 4 further comprising the mapping of emotive intensity numerical value into one or more word text describing the emotive intensity value in express language which would qualify an associated emotive state with the intensity value.

8. (original) A method as in claim 1 further comprising the scanning and tokenizing of the embedded emotive content in the communications.

9. (original) A method as in claim 1 further comprising parsing communications containing the emotive content using emotive grammar productions to tokenize the emotive content in textual communications.

10. (previously amended) A method of encoding emotive vectors, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author with associated text in electronic communications.

11. (original) The method in claim 10 further comprising structuring and synthesizing emotive parsers with productions exploiting emotive vectors encoded in textual datastreams.

12. (original) The method in claim 10 further comprising an emotive parser to tokenize emotive vectors into emotive components and emotive components to a set of face glyphs.

13. (original) The method in claim 12 further comprising a natural language parser to extract and tokenize emotive vector associated text into the parts of speech components.

14. (original) The method in claim 13 further comprising concatenating communication tokenized emotive components with grammatical string fragments and strings selected from the associated text into grammatical strings conveying an intended meaning of the communication.

15. (original) The method in claim 14 further comprising said face glyph set based on graphic rendering of reasonably representative emotive states and associated emotive intensities.

16. (previously amended) A computer program residing on a computer-readable media, said computer program communicating emotive content comprising emotive vectors, each emotive vector comprising an emotive state and an associated emotive intensity normalized to the author, with associated text embedded in electronic device communications.

17. (previously amended) A computer network comprising:

a plurality of computing devices connected by a network;

said computing devices which display graphical and textual output;

applications executing on the devices embedding emotive vectors which are representations of emotive states with associated author normalized emotive intensity;

assembling emotive content by associating emotive vectors with associated text in electronic communication;

encoding emotive content by preserving association of emotive vectors with associated text in the electronic communication;

transmitting the communication with emotive content to one or more receiver computing devices;

parsing communication bearing emotive content; and

mapping emotive vectors to face glyph representations from a set of face glyphs;

Such that communications encoded with emotive content facilitate exchange of precise emotive intelligence.

18. (original) A computer program residing on a computer-readable media, said computer program communicating over a computer network comprising:

a plurality of computing devices connected by a network;

said computing devices which display graphical and textual output;

computer-readable means for applications executing on the devices embedding emotive vectors which are representations of emotive states with associated author normalized emotive intensity;

computer-readable means for assembling emotive content by associating emotive vectors with associated text in electronic communication;

computer-readable means for encoding emotive content by preserving association of emotive vectors with associated text in the electronic communication;

computer-readable means for transmitting the communication with emotive content to one or more receiver computing devices;

computer-readable means for parsing communication bearing emotive content; and

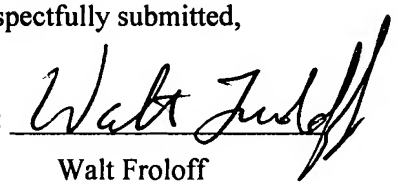
computer-readable means for mapping emotive vectors to face glyph representations from a set of face glyphs; and

computer-readable means for displaying communication of textual with associated face glyph emotive representations on said computing device displays;

whereby communications encoded with emotive content provide means of exchange of precise emotive intelligence.

If any matters can be resolved by telephone, applicant requests that the Patent and Trademark Office call the applicant at the telephone number listed below.

Respectfully submitted,

By: 

Walt Froloff

Walt Froloff
Inventor
273D Searidge Rd
Aptos, CA 95003
(831) 662-0505

Office Action Summary	Application No.	Applicant(s)	
	10/648,433	FROLOFF, WALT	
	Examiner	Art Unit	
	Cao (Kevin) Nguyen	2173	

-- *The MAILING DATE of this communication appears on the cover sheet with the correspondence address* --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *24 May 2007*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-18* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☒ Claim(s) *17 and 18* is/are allowed.

6)☒ Claim(s) *1-16* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Multimedia Tools and Applications 3, 105–125 (1996)

Generating and Manipulating Emotional Synthetic Speech on a Personal Computer

CAROLINE HENTON
Voice Processing Corporation, 1 Main Street, Cambridge, MA 02142

BRADLEY EDELMAN
Internet Products Group, Adobe Systems Inc., 1585 Charleston Road, P.O. Box 7900, Mountain View, CA 94039

Abstract. Against a background of incorporating a talking head into a role-playing simulator, enhancements are proposed for users of the simulator and of text-to-speech systems in general. The first is the ability to generate vocal emotion in synthetic speech using a limited number of prosodic parameters with a concatenative speech synthesizer. The second enhancement allows for vocal emotions to be included during the authoring of text for output by the text-to-speech system. Vocal emotions can be represented visually, and can be manipulated directly by the user. Applications such as training simulators that use synthetic speech can be made more 'human' by the addition of emotions. A graphical editor for specifying and directly manipulating the speech improves the authoring environment of these applications.

Keywords: emotions in synthetic speech, authoring training simulators, animated agents

1. Introduction

The central question we attempt to address in this paper is how to make an on-screen 'talking head' appear more human in its communication modes. To that end, we describe an authoring environment for producing vocal emotions in synthetic speech from parameters that can be manipulated using an intuitive visual interface.

At the outset we give the broad-based background to the need for such an authoring tool. Next, we review the literature and discuss limitations of current commercial systems that have any ability to simulate emotions in synthetic speech. We then describe how, using a limited number of prosodic controls, we can create vocal emotional affect in speech produced with a diphone-concatenative speech synthesizer. Specifically, the speech synthesizer is the one included in the text-to-speech (TTS) system named "MacinTalkPro 2®", first released on the Apple Macintosh Quadra 840 AV®[1] personal computer.

We give a detailed account of a user interface that represents speech parameters visually, and allows for their direct control. In contrast to previous techniques for authoring emotional synthetic speech, the approach presented here is embodied in a simplified format with a high level of abstraction. A user can easily predict how the text authored with the graphical editor will sound because of the explicit visual representation of vocal parameters.

For reasons of logic and clarity, the paper is divided into two parallel sections: the first is concerned with the speech controls, and the second focuses on a graphical user interface. This order of explanation is followed in all the sections: Section 3 presents and

critiques previous work in the speech domain. Sections 3.1–3.3 review the speech literature and provide an overview of how emotions have been simulated in previous work, and how they may now be integrated with greater facility into synthetic speech. The term 'vocal emotion' is clarified, showing how it is embodied in speech. A brief review and examples of how prosodic control is effected in a current commercial TTS system are given in Section 4. In Section 5 we outline our approach to simulating emotional affect, by using a limited number of acoustic prosodic controls. (See Glossary for all phonetic terms used in this paper). Section 6 focuses on the visual, graphic components. A sample implementation of the full authoring system is then presented. In summarizing our work, we indicate what we have found, and identify areas that require additional research. We conclude by exploring further possible applications of our findings.

2. Background

The work described here arose as part of a larger research endeavor that entailed participation by several groups of researchers; it is based in theories and methods employed in artificial intelligence (AI) and expert systems, computer graphics and multimedia, and text-to-speech (TTS) synthesis. The prime initiator lay in the development of a training, or role-playing simulator for needs analysis consultations. The role-playing simulator is used to teach students information gathering and communication skills, as detailed in Spohrer et al. [33] and further explained below.

In this training scenario, the student plays the role of a salesperson who attempts to gather information about the customer's computer networking needs. The student uses a menu-based language interface to interact with the simulated customer in a setting as similar as possible to a face-to-face meeting. The task being simulated is selling computer systems, which traditionally involves a technically qualified sales team (e.g. a systems engineer and a sales representative) meeting customers to perform a needs analysis consultation. The goal of the sales-team is to understand the customer's existing organization, systems, networking needs and special constraints; then to respond personably and rapidly to customers with a determination of relevant products and solutions, and with the ultimate goal of making a sale. The customer's responses are derived from a knowledge base and are presented by creating short digitized video Quicktime™ movies. The ultimate intent of the simulator is to provide a role-playing environment for a student systems engineer to experience contextualized actions and feedback, in which conversation is realistic and open-ended.

Although some concatenation techniques were used to string together frequently used phrases in the customer's turns in the dialogue, this approach to simulating the customer's audio-visual responses proved to be cumbersome for two reasons. First, the amount of disk space used grew rapidly and proportionately to the vocabulary of possible replies. The second, and more significant encumbrance, was that in order to expand or modify the vocabulary of customer replies, new video had to be shot and digitized. Such a need was expensive because it required re-creating the set, together with the additional time and effort of both the spokesmodel (the person 'speaking for', or modeling, the customer) and the technical staff involved in the filming/recording sessions. Further, even when significant efforts were made to avoid visual discrepancies, video shot during one session would rarely look identical to video shot during an differences in hair style, variations in l

Developments in computer graphic techniques that could be used to stret demonstrated in Patterson et al. [31] algorithm presented in Litwinowicz an as though it is talking (for fuller detail concatenative speech synthesizer for described in Henton [15]). From the neous projects, it was possible to conc to simulate a customer speaking on-s

To create a talking head, a photograp needed for eighty visually distinct dise as a Quicktime™ movie. In the Mac system, the synthetic speech, is pass manager provides interrupt informat duration. This information was used proper animation sequences, thus cre enhancements included the talking he given to a passage, and eye blinking. named 'MacHeadroom' see [18, 20]. meant that customer spoken respons simple text strings.

The attractiveness of such a synt 'script' can be stored as simple text the script is easy to modify or expan in a studio. A disadvantage of usin less natural, less human than those The interface presented in this pap synthetic customer replies. By pro speech synthesizer in an intuitive a some 'human-ness' into the syntheti

In short, the system integrates AI image warping and animation, toget authoring tool. It is a means to ma times when shooting more video fo

3. Speech components

3.1. *Previous work*

The ability to 'read aloud' text usi TTS) is not a recent invention. Th 50 years (for comprehensive revie

.1–3.3 review the speech literature
ılated in previous work, and how
ynthetic speech. The term 'vocal
:h. A brief review and examples of
TTS system are given in Section 4.
otional affect, by using a limited
)r all phonetic terms used in this
:nts. A sample implementation of
ng our work, we indicate what we
earch. We conclude by exploring

ıdeavor that entailed participation
d methods employed in artificial
d multimedia, and text-to-speech
ent of a training, or role-playing
ıying simulator is used to teach
as detailed in Spohrer et al. [33]

lesperson who attempts to gather
eeds. The student uses a menu-
stomer in a setting as similar as
:ed is selling computer systems,
m (e.g. a systems engineer and a
analysis consultation. The goal
ganization, systems, networking
and rapidly to customers with a
ı the ultimate goal of making a
:dge base and are presented by
ıltimate intent of the simulator
ystems engineer to experience
ı is realistic and open-ended.
string together frequently used
:h to simulating the customer's
reasons. First, the amount of
/ocabulary of possible replies.
order to expand or modify the
nd digitized. Such a need was
:h the additional time and effort
ıdeling, the customer) and the
Further, even when significant
uring one session would rarely

look identical to video shot during another session (witness, for example, minor physical differences in hair style, variations in lighting levels, etc.).

Developments in computer graphics made available high quality digital image warping techniques that could be used to stretch an image. By using a cross-mapping technique demonstrated in Patterson et al. [31] in conjunction with an image-warping ('morphing') algorithm presented in Litwinowicz and Williams [24] a photograph may be made to appear as though it is talking (for fuller details, see Henton and Litwinowicz [18]). Furthermore, a concatenative speech synthesizer for use on Macintosh computers had been developed (as described in Henton [15]). From the resources and techniques available in these simultaneous projects, it was possible to conceive of and create a 'talking head' that could be used to simulate a customer speaking on-screen.

To create a talking head, a photograph was chosen for a speaker. The animation sequences needed for eighty visually distinct disemes [16, 18] were recorded, pre-computed and stored as a Quicktime™ movie. In the Macintosh sound system, the output of the text-to-speech system, the synthetic speech, is passed to the speech manager to be spoken. The speech manager provides interrupt information about the next speech unit to be spoken and its duration. This information was used to set the appropriate playback rate and choose the proper animation sequences, thus creating the illusion of a talking head. Additional graphic enhancements included the talking head's eyebrows changing position based on the emotion given to a passage, and eye blinking. For further details on the talking head, internally code-named 'MacHeadroom' see [18, 20]. From the perspective of the simulation project, this meant that customer spoken responses could be generated in real time, from an input of simple text strings.

The attractiveness of such a synthesis of techniques is thus threefold: the customer's 'script' can be stored as simple text strings, which take up comparatively little disk space; the script is easy to modify or expand; the 'customer' does not need to re-create responses in a studio. A disadvantage of using a simulated speaker is that the synthetic replies are less natural, less human than those derived from the digitized movies of a spokesmodel. The interface presented in this paper was an effort to increase the effectiveness of the synthetic customer replies. By providing the author of the replies with control over the speech synthesizer in an intuitive and high-level manner, it was possible to re-introduce some 'human-ness' into the synthetic speech.

In short, the system integrates AI knowledge-based dialogue, text-to-speech synthesis, image warping and animation, together with a customizable text editor, to provide a novel authoring tool. It is a means to make rapid additions and alterations to a talking head at times when shooting more video footage of a human speaker would be impossible.

3. Speech components

3.1. Previous work

The ability to 'read aloud' text using synthetic speech (commonly called text-to-speech, TTS) is not a recent invention. The development of synthetic speech can be traced over 50 years (for comprehensive reviews see [1, 14, 15, 21, 30, 38]. Applications that include

speech (either short digitized files of real speech, or synthetic speech) on personal computers have been available for at least a decade, for example DECtalk®, and MacinTalk®. The parameters available for manipulation in DECtalk are described in detail by Klatt [21] and Klatt and Klatt [22]. Limitations of and constraints among the parameters in a parallel synthesizer such as DECtalk are critiqued by Stevens and Bickley [34]. In general, access to the speech parameters, and the ability to enhance the speech with emotional or other nuances, has been neither transparent nor friendly.

In the majority of its instantiations synthetic speech has been to date 'neutral' in tone, or, in the most parsimonious case, monotone. Synthetic speech has generally sounded disinterestedly dull, deficient in vocal emotionality. This deficiency is partly accounted for by the default intonation tunes in speech synthesizers which may be called 'wooden' or 'robotic'. Means may have existed to make the synthetic speech sound, for example, happy or angry, but research has been directed primarily towards maximizing intelligibility rather than including naturalness, or variety. Indeed, in the past two decades, some research ceased in TTS synthesis because it was believed that the largest problem, intelligibility, had been solved; for a critical commentary on this viewpoint see [35]. Previously published reports about the addition of emotional affect to synthesized speech have concentrated solely on parametric synthesizers and have used large numbers of acoustic parameters [3, 4, 28]. The study by Cahn [4] produced mixed results and remains inconclusive about "the perception of affect in speech" (p. 139).

In order to illustrate how synthetic speech can be provided with some emotional affect, by a relatively naive user, it is necessary to expand on three areas: (1) What is meant by vocal 'emotions'; (2) What acoustic correlates exist in speech for emotions; (3) Which, and how many, basic acoustic controls might be used to simulate emotions. The following sections address these questions. Details about how emotions are perceived in speech are not a concern here, since that issue is known to be an idiosyncratic and variable perceptual field [36]. There is tacit acknowledgement that the perception of synthesized emotions is not necessarily predictable and may not yet be a precise science.

3.2. *What is meant by 'vocal emotions'?*

Along a sliding scale of 'affect', voices may be heard to contain personalities, moods, and emotions. Personality was defined by Brown et al. [3] as "the characteristic emotional tone of a person over time". A mood may be considered to be a maintained attitude; whereas an emotion is a more sudden and more subtle response to a particular stimulus, lasting for seconds or minutes. The personality of a voice may therefore be regarded as its largest effect, and an emotion its smallest. The term 'vocal emotion' is used here to encompass the full range of affect in a voice.

Given the limitation of today's speech technology, and our limited understanding of factors involved in human speech production, it is currently impossible to re-create the full range of attributes of affect in the human voice in synthesized speech. However, many linguists and speech technologists argue that improvements in and the incorporation of these suprasegmental attributes are vital to the acceptability of synthetic speech, since these are precisely the components which extend synthetic speech beyond inhuman monotonicity,

and give to the speech its attitudinal ind
have underscored the need for the inte
an integral part of all speech, carrying
than the words themselves), emotion e

The literature on emotions indicates
describe rigorously [4, 28]. The interp
is only beginning to be understood.
cultural and semantic ambiguity. In
interpret emotions in recorded speech
different levels of sensitivity to emoti
vocal emotions are to some extent 'in

Different emotions have different le
in the literature about the scales along
scales are aggressiveness-pleasantnes
these scales, researchers generally rec
joy, sadness, fear, and disgust. Using
the five basic emotions may be used t
variants, e.g., grief, affection, sarcasm,
have not however found empirical sup
and identified than others, e.g. joy and
are anger and fear. Indifference is t
hardest to recognize.

Vocal emotion effects depend to som
voice quality differences [5, 23], intona
across languages. The findings descri
emotions in General American Englis

3.3. *What acoustic components in s*

Speech has two main components: v
and voice quality). The importance of
fact that children can understand emot
people who suffer from hearing-impai
tunes alone. Vocal components can cl
of the intended message as can the ve

Intonation is effected by suprasegm
speech segments. Voice quality (e.g.,
ing on the individual vocal tract; it a
affected by emotion are the pitch en
ing fundamental frequency, the pitch
overall speech rate, utterance timin
and intensity (loudness). Of these p
indicating emotion *per se*, but voice
emotions [5].

and give to the speech its attitudinal individuality [6, 10]. Murray and Arnott ([28], p. 1106) have underscored the need for the integration of these characteristics: "... as emotion is an integral part of all speech, carrying much of the information (and sometimes even more than the words themselves), emotion effects should be part of all synthetic speech".

The literature on emotions indicates that they are conceptually complex, and difficult to describe rigorously [4, 28]. The interplay between emotions, physiology and psychology is only beginning to be understood. Terms are used vaguely, and are plagued by cross-cultural and semantic ambiguity. In addition, the abilities of listeners to recognize and interpret emotions in recorded speech varies substantially. It appears that individuals have different levels of sensitivity to emotional stimuli. It has been found experimentally that vocal emotions are to some extent 'in the ear of the hearer'.

Different emotions have different levels of recognizability. There is, however, agreement in the literature about the scales along which emotions can be placed as discrete points: the scales are aggressiveness-pleasantness; interest-uninterest; authoritative-submissive. On these scales, researchers generally recognize and agree upon five 'basic' emotions: anger, joy, sadness, fear, and disgust. Using a 'palette' theory suggested by Scherer ([32], p. 43), the five basic emotions may be used to produce a larger number of (secondary) emotional variants, e.g., grief, affection, sarcasm, and surprise. The psychological bases of that model have not however found empirical support [29]. Some emotions are more readily expressed and identified than others, e.g. joy and sadness are easier to both express and identify than are anger and fear. Indifference is the emotion most easily recognized, and fear is the hardest to recognize.

Vocal emotion effects depend to some extent on language spoken (as well as age) and, like voice quality differences [5, 23], intonation [8] and grammar, are not necessarily transferable across languages. The findings described here are focused only on the synthesis of vocal emotions in General American English.

3.3. *What acoustic components in speech correlate with emotions?*

Speech has two main components: *verbal* (the words themselves), and *vocal* (intonation and voice quality). The importance of vocal components in speech may be indicated by the fact that children can understand emotions in speech before they can understand words, and people who suffer from hearing-impairment can still distinguish meaning from intonational tunes alone. Vocal components can clearly contribute as much to a listener's comprehension of the intended message as can the verbal, lexical components.

Intonation is effected by suprasegmental changes in the pitch, duration and amplitude of speech segments. Voice quality (e.g., nasal, breathy, or hoarse) is intrasegmental, depending on the individual vocal tract; it affects everything the speaker says. Voice parameters affected by emotion are the pitch envelope (as produced by a combination of the speaking fundamental frequency, the pitch range, the shape and timing of the pitch contour), overall speech rate, utterance timing (duration of segments and pauses), voice quality, and intensity (loudness). Of these parameters, it appears that pitch is more important in indicating emotion *per se*, but voice quality is more important in differentiating discrete emotions [5].

4. Current commercial TTS systems

Commercially available speech synthesizers use two distinct techniques: parametric and concatenative. Parametric speech synthesis is produced by mathematically manipulating individual acoustic parameters in time. The general methodology for controlling a parametric synthesizer is given in Allen et al. [1]. Concatenative speech synthesizers generate speech by linking pre-recorded speech segments to build syllables, words, or phrases. The size of the pre-recorded segments may vary from diphones, to demi-syllables, to whole words and phrases; see Henton [15] for further explanation of the two types of synthesis.

If computer memory and processing speed were unlimited, a possible method for creating vocal emotions might be to simply store words spoken by a human being in varying emotional ways. In the present state of the art, this approach is impractical. Rather than being stored, emotions have to be synthesized on-line and in real-time.

In parametric synthesizers (of which DECtalk is the most well-known and most successful), there may be as many as thirty basic acoustic controls available for altering pitch, duration and voice quality. These include, e.g. separate control of formants' values and bandwidths; pitch movements on, and duration of, individual segments; breathiness; smoothness; richness; assertiveness; etc. Precision of articulation of individual segments (e.g. fully released stops, degree of vowel reduction), which is controllable in DECtalk, can also contribute to the perception of emotions, such as tenderness and irony. These parameters may be manipulated to create voice personalities; DECtalk is supplied with nine different 'Voices' or personalities. It should be noted that intensity (volume) is not controllable within an utterance in DECtalk.

TTS systems also usually incorporate rules for the application of intonational attributes. In currently available systems, such as DECtalk and TrueVoice®, there is provision for the customization of the prosody and/or intonation of synthetic speech, generally using either high-level or low-level controls (see examples, below). However, these rule systems and controls are not well suited for authoring or editing emotional prose at a high level. The problem lies not only in the phonetically imprecise terminology, for example "baseline-pitch", but also in the difficulty of quantifying these terms. For example, if a user, untrained in phonetics or linguistics, wished to enter a stage play into a TTS system, to be read with synthetic speech, it would be unbearable (or, at the very least, challenging and overly time-consuming for the layperson) to have to choose numerical values for the various speech parameters in order to incorporate vocal emotion into each word spoken.

The high-level controls include text mark-up symbols, such as a pause indicator or pitch modifier. An example of such high-level text mark-up phonetic controls may be taken from the Digital Equipment Corporation DECtalk DTC03 Owner's Manual [9] where the input text string:

It's a mad mad mad mad world.

can have its prosody customized as follows:

It's a [/] mad [\] mad [/] mad [\] mad [^] world.

where [/] indicates pitch rise, and [\] indicates pitch fall.

Some synthesizers also provide the and pitch of phonetic symbols. These DECtalk:

[ow<1000>]

causes the sound [ow] (as in "over") to r (ms); while

[ow<,90>]

causes [ow] to receive its default duratic at the end; while

[ow<1000,90>]

causes [ow] to be 1000 ms long, and to

So, on the one hand, the disadvanta a very approximate effect and lack intu specification and the resulting vocal er impossible to achieve the desired inton control mechanism. On the other hand, t the intonational or vocal emotion specif expert analysis and testing (trial and erro in Hertz and milliseconds, by hand. T without considerable knowledge and tr

Most importantly, from our perspect commercial synthesizer described abov in scripts for TTS output.

5. Prosodic control in a concatenati

In diphone-concatenative speech synth control of individual acoustic features i the voice quality of the speaker, since speaker (who has their individual voice parameters for manipulating positions synthesizer. Secondly, precision of art in this type of synthesizer. It is none parameters listed in Table 1.

Details for using the commands lis Chapter 4 of *Inside Macintosh. Sounc* in Table 1, it is nevertheless possible

Some synthesizers also provide the user with direct control over the output duration and pitch of phonetic symbols. These are the low-level controls. Again, examples from DECtalk:

[ow<1000>]

causes the sound [ow] (as in "over") to receive a duration specification of 1000 milliseconds (ms); while

[ow<,90>]

causes [ow] to receive its default duration, but it will achieve a pitch value of 90 Hertz (Hz) at the end; while

[ow<1000,90>]

causes [ow] to be 1000 ms long, and to be 90 Hz at the end.

So, on the one hand, the disadvantage of the high-level controls is that they give only a very approximate effect and lack intuitiveness or direct connection between the control specification and the resulting vocal emotion of the synthetic speech. Further, it may be impossible to achieve the desired intonational or vocal emotion effect with such a coarse control mechanism. On the other hand, the disadvantage of the low-level controls is that even the intonational or vocal emotion specification for a single utterance can take many hours of expert analysis and testing (trial and error), including measuring and entering detailed values in Hertz and milliseconds, by hand. This is clearly not a task an average user can tackle without considerable knowledge and training in the various speech parameters available.

Most importantly, from our perspective, none of the studies cited in Section 3.1, nor the commercial synthesizer described above make any provision for direct authoring of emotion in scripts for TTS output.

5. Prosodic control in a concatenative synthesizer

In diphone-concatenative speech synthesizers, such as that included in MacinTalkPro 2, control of individual acoustic features is severely limited. Firstly, it is not possible to alter the voice quality of the speaker, since the speech is created from the recording of a live speaker (who has their individual voice quality) speaking in one (neutral) vocal mode, and parameters for manipulating positions of the vocal folds are not included in this type of synthesizer. Secondly, precision of articulation of individual segments is not controllable in this type of synthesizer. It is nonetheless possible in MacinTalkPro 2 to control the parameters listed in Table 1.

Details for using the commands listed in Table 1 in MacinTalkPro 2 are published in Chapter 4 of *Inside Macintosh. Sound* [19]. Although there are seven parameters listed in Table 1, it is nevertheless possible to produce a range of emotional affect using the

Table 1. Prosodic parameters available for control, with their associated commands, in MacinTalkPro 2.

Parameter	Speech synthesizer commands
1. Average speaking pitch	Baseline pitch (pbas)
2. Pitch range	Pitch modulation (pmod)
3. Speech rate	Speaking rate (rate)
4. Volume	Volume (volm)
5. Silence	Silence (slnc)
6. Pitch movements	Pitch rise (/), pitch fall (\)
7. Duration	Lengthen (>), shorten (<)

interplay of only *five* parameters—since Speech rate and Duration, and Pitch range and Pitch movements are, respectively, effected by the same acoustic controls.

Table 2, below, gives examples of some emotions which were defined, together with their associated vocal emotion values. These examples were chosen because they represent the emotions on which listeners most commonly reach perceptual consensus (cf. findings by Scherer [32], cited above). It should be remembered that these values were designed to apply to General American English, and to be specified for application to other d values shown are easily modifiable, to a user/listener perceptions.

The values (and underlying comments setting for a high-quality female voice i Table 2 would need to be altered. For e: male voice in MacinTalkPro 2 might u: specifying a lower, but more dynamic, general, neither volume nor speaking ra not need to be altered dramatically when [13]). As for determining values for oth voice, these values could merely chang the default specification. There is con variation is broad in the cross-dialect suprasegmental features, although Hen was employed relatively consistently a perception of emotions associated with values for other dialects and languages adjustable controls in MacinTalkPro 2. speech rate is 175 words per minute (v is 50–500 wpm.

The values shown in Table 2 are input set and calculations given in Chapter 4 c that the parameters pitch mean and pitc scale of semitones in the speech synthes frequency (see Glossary). The logarith in the range 0–100 for the convenienc are each represented on a logarithmic s On this basis, a pmod value of 6 will a pbas value of 26 than with 56. The therefore doubling of a volume value speech synthesizer used in MacinTalk

As detailed in Chapter 4 of *Inside M* line Pitch (pbas), Pitch Modulation Silence (slnc), may be applied at all phoneme, and allophone.

The following examples show the re portions of text. The first scenario s text-to-speech system and using the In this scene, the portions of text in it while the rest of the text indicates the the speech synthesizer parameters; a comments added for clarification her

Table 2. Examples of some vocal emotions defined according to a restricted set of prosodic values. N.B. These values were designed to apply to a female voice speaking General American English, only.

Emotion	Pitch mean/range (pbas)/(pmod)	Volume (volm)	Speaking Rate (rate)
Default (normal)	56; 6 (Neutral and narrow)	0.5 (Neutral)	175 Neutral
Angry1 (threat)	35; 18 (Low and narrow)	0.3 (Low)	125 (Slow)
Angry2 (frustration)	80; 28 (High and wide)	0.7 (High)	230 (Fast)
Happy (medium)	65; 30 (Neutral and wide)	0.6 (Neutral)	185
Curious	48; 18 (Neutral and narrow)	0.8 (High)	220 (Fast)
Sad	40; 18 (Low and narrow)	0.2 (Low)	130 (Slow)
Emphasis	55; 2 (Neutral and narrow)	0.8 (High)	120 (Slow)
Bored (medium)	45; 8 (Neutral and narrow)	0.35 (Low)	195
Aggressive	50; 9 (Neutral and narrow)	0.75 (High)	275 (Fast)
Tired	30; 25 (Low and neutral)	0.35 (Low)	130 (Slow)
Disinterested	55; 5 (Neutral)	0.5 (Neutral)	170 (Neutral)

ciated commands, in MacinTalkPro 2.

thesizer commands
)itch (pbas)
lulation (pmod)
rate (rate)
'olm)
lnc)
(/), pitch fall (\)
(>), shorten (<)

:nd Duration, and Pitch range and
: acoustic controls.
:h were defined, together with their
chosen because they represent the
:eptual consensus (cf. findings by
hat these values were designed to

stricted set of prosodic values. N.B. These
erican English, only.

ıme m)	Speaking Rate (rate)
5	175
ıral)	Neutral
3	125
w)	(Slow)
r	230
h)	(Fast)
:	185
ral)	
	220
h)	(Fast)
	130
')	(Slow)
	120
ı)	(Slow)
	195
)	
	275
)	(Fast)
	130
)	(Slow)
	170
l)	(Neutral)

apply to General American English, and the user would need different vocal emotion values to be specified for application to other dialects and languages. Nevertheless, the particular values shown are easily modifiable, to allow for differences in cultural interpretations and user/listener perceptions.

The values (and underlying comments) in Table 2 are relative to the default neutral speech setting for a high-quality female voice in MacinTalkPro 2. For a male voice, the values in Table 2 would need to be altered. For example, the default specification for a high-quality male voice in MacinTalkPro 2 might use a pitch mean of 43 and a pitch range of 8 (thus specifying a lower, but more dynamic, range than the female voice of 56; 6). However, in general, neither volume nor speaking rate is sex-specific, and, as such, these values would not need to be altered dramatically when changing the sex of the speaking voice (cf. Henton [13]). As for determining values for other vocal emotions when changing to a male speaking voice, these values could merely change as the female voice specifications do, relative to the default specification. There is considerable agreement in the phonetic literature that variation is broad in the cross-dialect and cross-language use of prosodic patterns and suprasegmental features, although Henton [12, 17] found that pitch range and dynamism was employed relatively consistently across sexes and across dialects. The cross-cultural perception of emotions associated with those patterns is even more variable. Appropriate values for other dialects and languages would have to be established empirically using the adjustable controls in MacinTalkPro 2. It should be noted that in MacinTalkPro 2 the default speech rate is 175 words per minute (wpm) whereas a realistic human speaking rate range is 50–500 wpm.

The values shown in Table 2 are input to the speech synthesizer, according to the command set and calculations given in Chapter 4 of *Inside Macintosh. Sound* [19]. We need to point out that the parameters pitch mean and pitch range are represented acoustically in a logarithmic scale of semitones in the speech synthesizer, where 12 semitones correspond to a doubling in frequency (see Glossary). The logarithmic values are converted to a linear scale of integers in the range 0–100 for the convenience of the user. Because pitch mean and pitch range are each represented on a logarithmic scale, the interaction between them is quite sensitive. On this basis, a pmod value of 6 will produce a markedly different perceptual result with a pbas value of 26 than with 56. The range for volume, on the other hand, is linear and therefore doubling of a volume value results in a doubling of the output volume from the speech synthesizer used in MacinTalkPro 2.

As detailed in Chapter 4 of *Inside Macintosh. Sound* [19], prosodic commands for Baseline Pitch (pbas), Pitch Modulation (pmod), Speaking Rate (rate), Volume (volm), and Silence (slnc), may be applied at all levels of text, i.e., passage, sentence, phrase, word, phoneme, and allophone.

The following examples show the results of applying different vocal emotions to different portions of text. The first scenario shows the result of merely inputting the text into the text-to-speech system and using the default vocal emotion parameters for female voices. In this scene, the portions of text in italics indicate speech by the car repair-shop employee while the rest of the text indicates the car owner. The portions in double brackets indicate the speech synthesizer parameters; and the portions of text in single brackets are merely comments added for clarification here.

1. [Default] [[pbas 56; pmod 6; rate 175; volm 0.5]] Is my car ready? *Sorry, we're closing for the weekend.* What? I was promised it would be done today. I want to know what you're going to do to provide me with transportation for the weekend!

With only the default prosodic values in place, MacinTalkPro 2 could play this scenario through a loudspeaker, however, it would be hard to distinguish the two speakers in the conversation, and the interchange might sound somewhat robotic owing to the lack of vocal emotion. After the application of vocal emotion parameters (either through use of the graphical user interface, direct textual insertion, or other automatic means of applying the defined vocal emotion parameters), the text might look like the following:

2. [Default] [[pbas 56; pmod 6; rate 175; volm 0.5]] Is my car ready? [Disinterested] [[pbas 55; pmod 5; rate 170; volm 0.5]] *Sorry, we're closing for the weekend.* [Angry1] [[pbas 35; pmod 18; rate 125; volm 0.3]] What? I was promised it would be done today. [Angry2] [[pbas 80; pmod 28; rate 230; volm 0.7]] I want to know what you're going to do to provide me with transportation for the weekend!

This second scenario thus provides the speech synthesizer with parameters that will result in the output having vocal emotion. It should be noted that two varieties of 'Anger' are suggested; this emotion has been shown to have two distinct manifestations in speech Frick [11]. The first ('Angry1') may be heard as 'cold' anger, a form of controlled threat; the second ('Angry2') is 'hot' anger, being louder, faster, more dynamic and uncontrolled. The addition of these vocal emotions is likely to provide the listener with much greater content than merely hearing the words spoken in an emotionless manner.

Individual words within a passage can receive only one type of modification, specifically where additional emphasis [[emph]] on a single (following) word is achieved by a rise in the pitch and a lengthening of the vowels:

```
[[pbas 56; pmod 6; rate 175; volm 0.5]]
This is a [[emph +]] beautiful [[slnc 30]] morning. [[rate 140; volm 0.4]]
The sun is piercing the sky between [[rate 150; volm 0.6]]
black [[rset]] clouds that cling to the Santa Cruz mountains' crest.
```

Both [[emph]] and [[slnc]] apply only to the following word string, and do not require resetting, or toggling off.

MacinTalkPro 2 also gives the user access to phonemes, the minimal contrastive units of speech. The exact specification of the phonemes used for General American English is not needed here. Modifications to individual phonemes within a passage can be achieved by first entering the phonemic Input Mode [[inpt PHON]] and then adding prosodic inflection controls to the basic phoneme symbols, as illustrated in the example below for the word "anticipation" in the phrase "Anticipation is all":

```
[[inpt PHON]]/2AEn = t2IH = sIX = p1 >/EY = S/IXn[[inpt TEXT]] is all.
```

The pronunciation of the word "ant than normal, because of the rising pitc the increased length (>) of the penulti by the equals sign (=).

Modifications to allophones (see *In* used to achieve the lowest level effect entering the allophonic Input Mode, numerical values for duration (D) and Bob". below:

```
[[xtnd gala inpt ALLO]]
h [D 90][P 120 : 50]
AY[D 274][P 227 : 5, 213 : 30
;
b[D 140][P 120 : 50]
AA [D 420][P 88 : 5, 85 : 30, 1
b - [D 30][P 120 : 50]
.
[[inpt TEXT]]
```

For Duration (D), the integer is m absolute pitch target (in Hertz), or th reached, and the second value gives reached. For example, the final 'b-' ha is reached at 50% of its total duration allophones and associated prosodic va Similarly, the semicolon word separa acoustic effect; they are included to hel at the allophonic level, whereby the te certain time into the sound, and the rel volume (0–100); see *Inside Macintos*

It is possible to experiment with sy emotional connotation. Inflection Cc provide more exaggerated, cumulative the speech synthesizer, and on its perc

6. Visual speech parameters

As illustrated above, terms used in sp well suited for authoring emotional p terminology, but also in the difficulty numerical values for each of several each word spoken would be very tires

ls my car ready? *Sorry, we're closing*
be done today. I want to know what
on for the weekend!

cinTalkPro 2 could play this scenario
› distinguish the two speakers in the
'hat robotic owing to the lack of vocal
ırameters (either through use of the
her automatic means of applying the
k like the following:

]] Is my car ready? [Disinterested]
re closing for the weekend. [Angry1]
vas promised it would be done today.
I want to know what you're going to
ınd!

ːsizer with parameters that will result
ed that two varieties of 'Anger' are
stinct manifestations in speech Frick
ger, a form of controlled threat; the
nore dynamic and uncontrolled. The
ıe listener with much greater content
ːss manner.
ɔne type of modification, specifically
ɔwing) word is achieved by a rise in

[[rate 140; volm 0.4]]
n 0.6]]
ɔuntains' crest.

ing word string, and do not require

ıes, the minimal contrastive units of
ːor General American English is not
'ithin a passage can be achieved by
and then adding prosodic inflection
in the example below for the word

' = S/IXn[[inpt TEXT]] is all.

The pronunciation of the word "anticipation" could be perceived as being more excited than normal, because of the rising pitch (/) on the first, penultimate and last syllables, and the increased length (>) of the penultimate syllable. In this notation, syllables are divided by the equals sign (=).

Modifications to allophones (see *Inside Macintosh. Sound*, [19], pp. 4–33), which are used to achieve the lowest level effects on the pronunciation of a string, are made by first entering the allophonic Input Mode, [[xtnd gala inpt ALLO]], and then adding prosodic numerical values for duration (D) and Pitch (P), as illustrated in the example phrase "Hi Bob". below:

```
[[xtnd gala inpt ALLO]]
h [D 90][P 120 : 50]
AY[D 274][P 227 : 5, 213 : 30, 196 : 55, 136 : 80]
;
b[D 140][P 120 : 50]
AA [D 420][P 88 : 5, 85 : 30, 119 : 55, 151 : 80]
b - [D 30][P 120 : 50]
.
[[inpt TEXT]]
```

For Duration (D), the integer is milliseconds. For Pitch (P), the first value is for the absolute pitch target (in Hertz), or the relative target (relative pitch number 1–99) to be reached, and the second value gives the time into the segment that the target should be reached. For example, the final 'b-' has a duration of 30 milliseconds, and a pitch of 120 Hz is reached at 50% of its total duration, or half-way into the sound. In the example above, allophones and associated prosodic values are listed by line-by-line for ease of readability. Similarly, the semicolon word separator and the final period are optional. Neither has an acoustic effect; they are included to help readers. A volume control can also be implemented at the allophonic level, whereby the target volume is given as an integer to be reached at a certain time into the sound, and the relative volume represents a percentage of the maximum volume (0–100): see *Inside Macintosh. Sound* ([19], pp. 4–29).

It is possible to experiment with synergistic combinations of settings to achieve a given emotional connotation. Inflection Control symbols (/, \, <, >) may be concatenated to provide more exaggerated, cumulative effects. The specific nature of the effect depends on the speech synthesizer, and on its perception by the listener.

6. Visual speech parameters

As illustrated above, terms used in speech synthesis and existing prosodic controls are not well suited for authoring emotional prose at a high level. The problem lies not only in the terminology, but also in the difficulty of quantifying these terms. To reiterate: choosing numerical values for each of several speech parameters to incorporate vocal emotion into each word spoken would be very tiresome. A more intuitive and faster approach is needed.

Of course, other graphical interfaces for modification of sound currently exist. For example, commercial products such as SoundEdit®, by Farallon Computing, Inc., provide for manipulation of raw sound waveforms. However, SoundEdit does not provide for direct user manipulation of the waveform (instead, the portion of the waveform to be modified is selected and then a menu selection is made for the particular modification desired). Manipulation of raw waveforms does not provide a clear intuitive means to specify vocal emotion in synthetic speech because of the lack of clear connection between the displayed waveform and the desired vocal emotion. Simply put, by looking at a waveform of human speech, an acoustically naive user cannot easily ascertain how it (or modifications to it) will sound when played through a loudspeaker, particularly if the user is attempting to provide some sort of vocal emotion to the speech.

We will now present a graphical user interface which gives the user of our speech synthesizer a way to harness speech parameters. The interface was designed so that the user does not need to have a knowledge or understanding of the underlying speech synthesizer. Instead the user is provided with a visual representation and direct manipulation. The interface builds upon the elements of soundwave editors such as SoundEdit mentioned above. However, the interface we suggest is extended in new ways which allow speech to be considered at a higher, and more understandable, level than a waveform. In addition, not only can the amplitude and temporal attributes of the sound be edited, but high level effects such as emotion can also be introduced.

Our interface allows the user to visually represent and to control the following vocal characteristics through direct manipulation: volume, duration, pitch variation. By combining the acoustic parameters, the user can introduce vocal emotion. The desired implementation takes the form of a standard text-editing system which provides the additional functionality we describe.

Figure 1 is a simplified block diagram of the stages involved in applying emotion to synthetic speech using our graphical interface.

Figure 1. Simplified flow diagram of stages interface.

6.1. *Visual volume and duration*

As may be seen in a sound waveform editor, the control of volume and duration takes advantage of the two natural spatial axes of a computer display; volume is the vertical axis, duration the horizontal axis. By single clicking on a word in the text to be output by the text-to-speech system, that word is selected and available for manipulation. Three sizing grips are presented: one for volume only, one for duration only, and one which allows both volume and duration to be manipulated simultaneously. The word is simply stretched along the axes. The taller a word becomes, the greater volume it will have; likewise, the wider a word becomes, the greater duration it will have. The manipulation is straightforward, and the resulting visual feedback and representation allows the user to understand volume and duration content at a glance. This direct mapping is a great improvement over embedded commands such as [[volm 0.7]] or [[rate 180]]. An analogy could be drawn to the immediate clarity of a graph compared with the table of numerical values it plots. One is obvious, while the other is difficult to interpret. Figure 2 illustrates this notion. The original text is shown, followed by a series of manipulations required to create the resulting text.

6.2. *Visual emotion*

Emotion can also be added to the te
Colors are used to associate an em
requires a computer with a multi-co
as they seem appropriate for the en
implications. For example, in some
others yellow may be perceived as
color red will represent angry, and y
cat speaking the sentence "Pete's go
would highlight it in the manner stan
from a range of emotions. The sel
being, of course, independent from
'Angry' reply to his cat would be sl
with the color black and is the defa

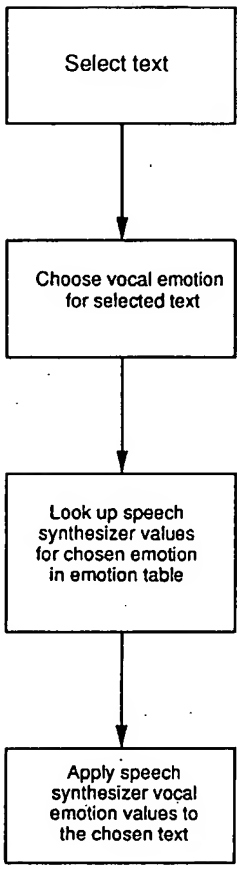


Figure 1. Simplified flow diagram of stages involved in applying emotion to synthetic speech using our graphical interface.

6.2. *Visual emotion*

Emotion can also be added to the text using direct manipulation and visual representation. Colors are used to associate an emotion with a word. This component of our interface requires a computer with a multi-color display. Colors may be chosen by the implementor as they seem appropriate for the emotions in question, and to allow for differing cultural implications. For example, in some cultures, yellow may be perceived as happy, while in others yellow may be perceived as angry. However, for the sake of illustration here, the color red will represent angry, and yellow will represent happy. Accordingly, imagine Pete's cat speaking the sentence "Pete's goldfish was delicious". The user authoring this sentence would highlight it in the manner standard in modern text editing systems, and select 'Happy' from a range of emotions. The selected text would then turn yellow, the change in color being, of course, independent from its other attributes (its volume and duration). Pete's 'Angry' reply to his cat would be shown in red. An emotion called 'Normal' is associated with the color black and is the default. This concept is illustrated in figure 3.

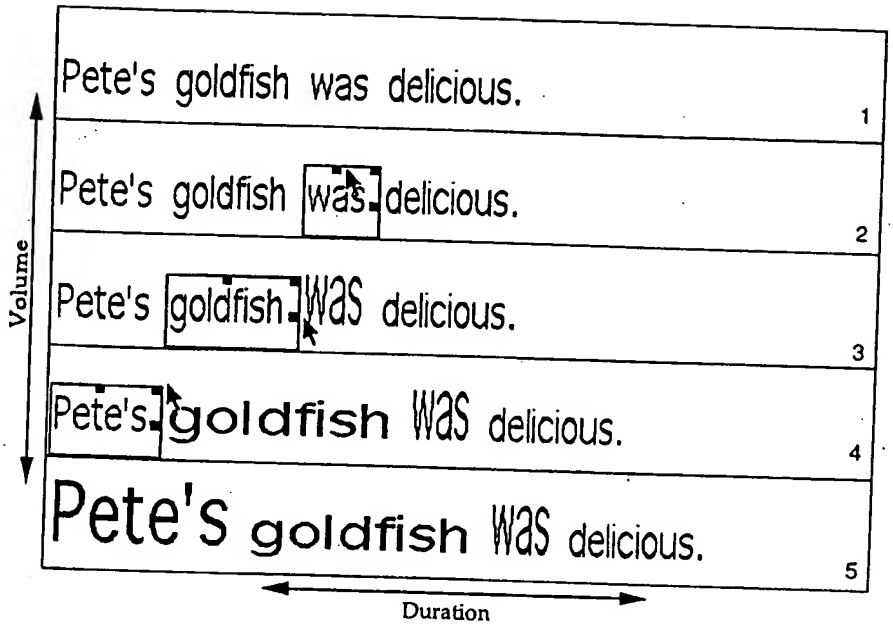


Figure 2. Original text (line 1) can have a series of manipulations applied to it to create the resulting text (line 5). Three sizing grips are provided: one for volume only (line 2), one for duration only (line 3), and one which allows both volume and duration to be manipulated simultaneously (line 4). The selected word is stretched using a pointing device (the arrow shaped mouse cursor of the Macintosh is shown) along the axes, labeled here 'Duration' on the x-axis and 'Volume' on the y-axis. Line numbers are included for clarity only here, and do not appear on the graphical editor screen.

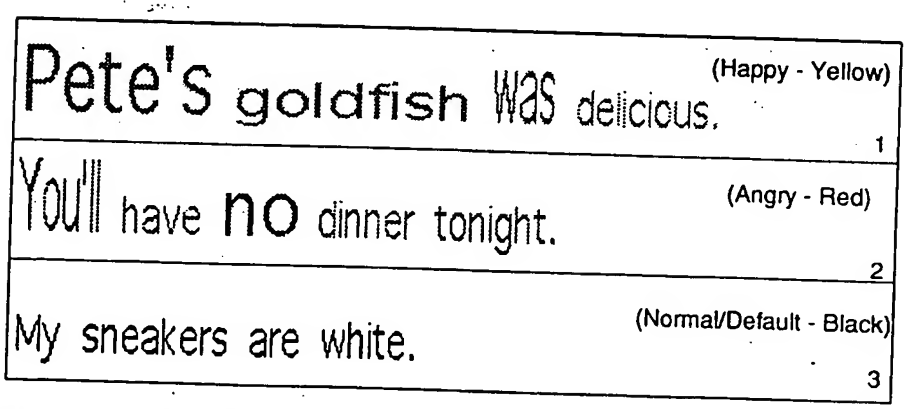


Figure 3. In line 1, text is highlighted to select 'Happy' from a range of emotions. The selected text turns yellow; the change in color is independent from its other attributes (its volume and duration). In line 2, the 'Angry' reply appears in red. In line 3, a default emotion ('Normal') is associated with the color black. Line numbers, emotions and colors are included for clarity only here; they do not appear on the graphical editor screen.

Figure 4. Pitch controls may be inserted into syllable on which the pitch change will occu pitch-fall by a left-to-right downward slope.

Again, a direct and intuitive visual teristic which has proved difficult to r

6.3. *Visual pitch variation*

Our graphical interface also allows f inserted by dropping 'pitch marks' in the pitch change will occur. A rise in drop in pitch by a left-to-right downw

6.4. *Mapping between the visual an*

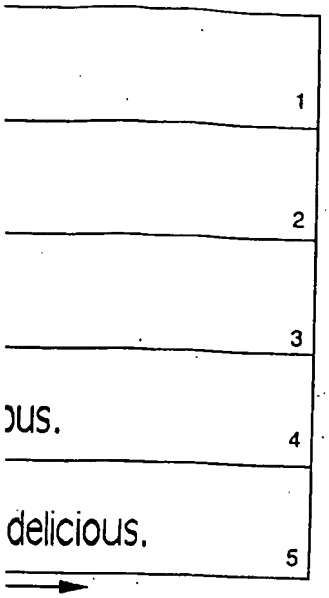
In this environment, the mapping of v formation. Visually, the font is being y% of its normal size vertically. An editor through a user preference dialo allows for sufficient dynamic range a volume settings and speech rate setti inversely proportional to duration) are is performed by the interface during t

The mapping of emotion is less st speech synthesizer used, MacinTalkPr for each prosodic parameter for each is designated for an emotion, the map matter of table look-up. We used the

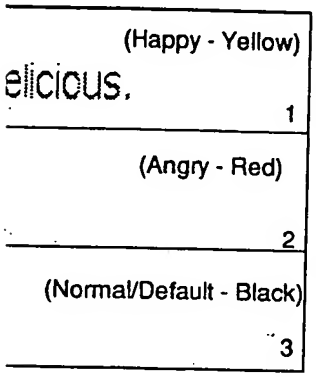
The translation of pitch variation is provided by the speech synthesizer. specified value <n> in [[pbas + <n>

7. Sample implementation

As stated above, the interface is simpl text editor from the simple (e.g. Teac be extended to support our interface. F editor. A screen shot of that editor ap



ed to it to create the resulting text (line 5). r duration only (line 3), and one which 4). The selected word is stretched using a n) along the axes, labeled here 'Duration' r clarity only here, and do not appear on



notions. The selected text turns yellow; l duration). In line 2, the 'Angry' reply ie color black. Line numbers, emotions iphical editor screen.

Figure 4. Pitch controls may be inserted into text by dropping 'pitch marks' into the document above the desired syllable on which the pitch change will occur. A pitch-rise is represented by a left-to-right upward slope, a pitch-fall by a left-to-right downward slope.

Again, a direct and intuitive visual representation is offered for a complex vocal characteristic which has proved difficult to represent and understand quantitatively.

6.3. *Visual pitch variation*

Our graphical interface also allows for the control of changes in pitch. Pitch controls are inserted by dropping 'pitch marks' into the document above the desired syllable on which the pitch change will occur. A rise in pitch is represented by a left-to-right upward slope, a drop in pitch by a left-to-right downward slope. This concept is illustrated in figure 4.

6.4. *Mapping between the visual and parametric representations*

In this environment, the mapping of volume and duration is a straightforward linear transformation. Visually, the font is being displayed at x% of its normal size horizontally, and y% of its normal size vertically. An allowable range of percentages is established by the editor through a user preference dialog, (for example between 50 and 200 percent), which allows for sufficient dynamic range and a manageable display. Corresponding ranges of volume settings and speech rate settings (for our simplified purposes here, speech rate is inversely proportional to duration) are established and the appropriate linear normalization is performed by the interface during the translation.

The mapping of emotion is less straightforward and more subjective. In the particular speech synthesizer used, MacinTalkPro 2, it is possible to choose experimentally the values for each prosodic parameter for each of the emotions desired. Once a set of parameters is designated for an emotion, the mapping between color and parameterization becomes a matter of table look-up. We used the values in Table 2 in our implementation.

The translation of pitch variation is a straightforward mapping to the appropriate controls provided by the speech synthesizer. In our case a rising pitch line is mapped to a user-specified value <n> in [[pbas + <n>]].

7. Sample implementation

As stated above, the interface is simply an extension to a standard text editing system. Any text editor from the simple (e.g. TeachText®) to the monolithic (Microsoft Word®) could be extended to support our interface. For our purposes, we implemented our own basic text editor. A screen shot of that editor appears in figure 5.

Figure 5. Screen shot of our editor. The buttons bearing the names of different emotions, with their associated colors indicated above them in parentheses, are used to change the vocal emotion of the text.

As illustrated above, individual words may be selected. Words can be 'stretched' along both the vertical and horizontal axes, to scale both volume and duration respectively. The buttons bearing the names of different emotions can be used to change the emotion of the currently-selected word or words.

As in any standard text editor, words can be inserted, deleted, cut, copied, pasted, etc. The intent of the text editor interface extension is simply to allow for the introduction of vocal emotions into the prose while preserving a familiar and well proven text editing environment.

8. Conclusion

Recently Vitale ([37], p. 25) made the following prediction: "Speech synthesizers of the future will offer a range of emotional parameters which will provide users with the ability to convey various emotions by allowing the prosodics to match the semantics of the utterance. A user will be able to produce a sente: fervor rather than boredom". In the w important strides towards fulfilling th

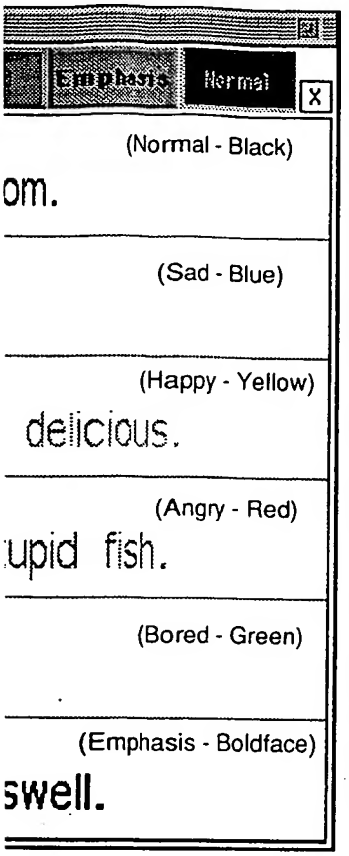
This synergistic work contains sev simulated dialogs. Linguistic/acous emotions to the synthetic speech. I concatenative speech synthesis syste dividual manipulation to simulate vo and how, is not previously reported, a lines we offer for the direct manipula to the best of our knowledge, a new f ing system provides an expeditious and alterations to the speech and rel

Additional research is required int impact on understanding or tolerab application such as MacHeadroom. large amounts of synthetic speech in the speech (cf. criticisms of the intr p. 22) and by Cowley and Jones ([7 TTS systems are not currently impre that has more or less reached asymp of voices. The latter are of particula Vitale [37], pp. 20–23). Furthermo to on-screen animated characters w or modify a single synthetic voice n Jones' [7] findings about users ratin

Judgements on the comparative the presence of MacHeadroom sho psychology that investigates poten mation. For example, Massaro and speech perception for a considerab on the McGurk effect, on speech-r languages. The talking head used b (known as 'Baldy') and the synthe be instructive to explore the comp more human head, namely MacHea inTalkPro 2, in these types of exper to establish any differences in the r the presence or absence of a MacH

Further possible applications of puter agents, which could be visu personalized using the custom-m sizer. A talking head might enha read over the telephone or on-scr

1) (Boldface) (Black)



lifferent emotions, with their associated emotion of the text.

Words can be 'stretched' along and duration respectively. The ed to change the emotion of the

ed, cut, copied, pasted, etc. The ow for the introduction of vocal roven text editing environment.

n: "Speech synthesizers of the provide users with the ability to the semantics of the utterance.

A user will be able to produce a sentence such as "*This is exciting technology*!" and convey fervor rather than boredom". In the work described here, we consider we have made several important strides towards fulfilling that prediction.

This synergistic work contains several novel concepts. It integrates synthetic speech into simulated dialogs. Linguistic/acoustic theory is used to suggest possibilities for adding emotions to the synthetic speech. Regarding the method of speech synthesis used, any concatenative speech synthesis system will have a set of prosodic controls available for individual manipulation to simulate vocal emotions or personalities; which controls are used, and how, is not previously reported, as far as we are able to determine. In addition, the guidelines we offer for the direct manipulation and visual representation of emotional speech are, to the best of our knowledge, a new facility in application authoring. Ultimately, the authoring system provides an expeditious prototyping tool and a means to make rapid additions and alterations to the speech and related facial expressions of an on-screen talking head.

Additional research is required into the perceived increase in naturalness, and the general impact on understanding or tolerability of synthetic speech from its embodiment in an application such as MacHeadroom. Listeners currently find it very unpleasant to listen to large amounts of synthetic speech in training applications, regardless of the intelligibility of the speech (cf. criticisms of the intrusiveness and quality of 'machine voice' by Baber ([2] p. 22) and by Cowley and Jones ([7], p. 149). According to Tatham ([35], p. 35), users of TTS systems are not currently impressed by synthetic speech; they want intelligibility (and that has more or less reached asymptote) but they also want naturalness and a wider range of voices. The latter are of particular concern to persons with diabilities (for a summary see Vitale [37], pp. 20–23). Furthermore, listeners have been observed to respond differently to on-screen animated characters when the synthetic voice changes. The ability to enhance or modify a single synthetic voice may therefore increase user acceptance (cf. Cowley and Jones' [7] findings about users ratings of the task-appropriateness of synthetic voices).

Judgements on the comparative qualitative experience of listening to TTS with/without the presence of MacHeadroom should also be obtained. There is a large body of work in psychology that investigates potential trade-offs in perceiving visual and auditory information. For example, Massaro and colleagues have conducted research into audio-visual speech perception for a considerable time (see *inter alia* [25–27]). Their focus has been on the McGurk effect, on speech-reading and on the transferability of such effects across languages. The talking head used by Massaro et al. is a Parkes geometric articulatory frame (known as 'Baldy') and the synthesizer is a parametric one, similar to DECtalk. It would be instructive to explore the comparative effectiveness and/or acceptability of a different, more human head, namely MacHeadroom, and a different type of synthesizer, namely MacinTalkPro 2, in these types of experiments. Similarly, perception tests need to be conducted to establish any differences in the reaction time taken to respond to instructions given with the presence or absence of a MacHeadroom-like on-screen agent.

Further possible applications of our findings include the more widespread use of computer agents, which could be visually personalized from a still photograph, and vocally personalized using the custom-made text editor in combination with a speech synthesizer. A talking head might enhance the spoken delivery of electronic mail, and faxes read over the telephone or on-screen at the desktop. It could also be incorporated into

computer-telephony-interfaced (CTI) applications such as automated receptionists that manage an owner's schedule and can be programmed to prioritize, sort and announce telephonic access to the owner of the system. It is also possible to envisage many educational uses, such as assisting in the acquisition of reading skills, and first or second language-learning.

As stated at the beginning of the paper, the longer-term objective of this work was to provide an interface to the role of a simulated customer in a training simulator. Some potential advantages of learning with simulators are listed by Spohrer et al. [33]: "increased time on task, on demand learning, safety, supportiveness, and transparency". We have made a convincing attempt to overcome some of the difficulties in using bimodal text-to-speech synthesis. By integrating MacHeadroom, a talking head, into the training simulation and designing a tool for authoring text spoken synthetically, we consider we have added a significant real-time, computationally low-cost enhancement in human-computer communication, while simultaneously reducing computing bandwidth and development effort in the role-playing simulator.

Acknowledgments

The authors are grateful for the valuable insights, constructive criticism, collaborative research and implementing help provided by Randy Gard, Arthur James, Peter Litwinowicz, Scott Meredith, James Spohrer, and Lance Williams, all in the Advanced Technology Group at Apple Computer, Inc., Cupertino, CA 95014.

Note

1. MacinTalkPro 2® and Macintosh Quadra 840 AV® are registered trademarks of Apple Computer, Inc.

Glossary

Terms which are cross-referenced in the glossary appear in bold print.

Allophone: a context-dependent variant of a **phoneme**. For example, the [*t*] sound in 'train' is different from the [*t*] sound in 'stain'. Both [*t*]s are allophones of the phoneme /*t*/. Allophones do not change the meaning of a word, they are all very similar to one another, but they appear in different phonetic contexts.

Concatenative synthesis: generates speech by linking pre-recorded speech segments to build syllables, words, or phrases. The size of the pre-recorded segments may vary from diphones, to demi-syllables, to whole words.

Duration: the length of a speech unit (word, syllable, **phoneme, allophone**). See **Length**.

General American English: a variety of American English that has no strong regional accent, and is typified by Californian, or West Coast American English.

Intonation: the pattern of **pitch** changes which occur during a **phrase** or **sentence**. E.g. the statement "You are reading" and the question "You are reading?" will have different intonation patterns, or tunes.

Length: the duration of a sound or sequence of sounds, usually measured in milliseconds (ms). For example, the vowel in 'cart' has greater intrinsic duration (is intrinsically longer) than the vowel in 'cat', when both words are spoken at the same speaking **rate**.

Phone: the phonetic term used for instantiations of real speech sounds, i.e., concrete realizations of **phonemes**.

Phoneme: any sound that can change the mea
all the pronunciations of similar context-dep
encode the transition from written letters to a
appropriate sound segments (**allophones**).

Pitch: the perceived property of a sound or sen
Pitch is the perceptual correlate of the fundar
movements are effected by falling, rising, an
many high falling pitch contours, and bored

Pitch range: the variation around the average
intonational contours. Pitch range has a med

Prosody: a collective term used for the varia
together with the variations in the rate of sp

Rate: the speed at which speech is uttered, u
words per minute (wpm). Allegro speech is
perception of the speech **style**.

Semitone: a pitch interval halfway between tv
scale is non-linear and interval-preserving. T
given in *Inside Macintosh. Sound* (1994, pp.

Speaking fundamental frequency: the avera
'baseline pitch'.

Speech style: the way in which an individual s
etc. Speech style will also be affected by the
styles, and how the speaker feels about what

Stop consonant: any sound produced by a total
American English, that appear initially in th

Suprasegmental: a phonetic effect that is not l
and which extends over an entire word, phr
suprasegmental elements of speech.

Vocal cords: the two folds of muscle, located i
vibrating, they may assume a range of positi
to fully open as in quiet breathing. Voiceles
pitch and in **voice quality** are produced by

Voice quality: a speaker-dependent characteris
are most quickly identified. Such factors as a
speaking situation will affect voice quality;
from New York City are thought to have m
nervous speaker may have a breathy and tre

Volume: the overall amplitude or loudness at

References

1. J. Allen, M.S. Hunnicutt, and D. Klatt, F
Press: Cambridge, 1987.
2. C. Baber, "Speech output," in Interactive
Francis: London, 1993, pp. 21–24.
3. B.L. Brown, W.J. Strong, and A.C. Ren
manipulations of rate, mean fundamental
personality from speech," Journal of the A
4. J.E. Cahn, "Generating expression in synt
sachusetts Institute of Technology, Cambr
5. R. Carlson, B. Granström, and I. Karlsson
Communication, Vol. 10, pp. 481–489, 19
6. R. Collier, "Multi-language intonation syr

ıch as automated receptionists that
to prioritize, sort and announce tele-
ssible to envisage many educational
skills, and first or second language-

-term objective of this work was to
mer in a training simulator. Some
listed by Spohrer et al. [33]: "in-
portiveness, and transparency". We
of the difficulties in using bimodal
m, a talking head, into the training
n synthetically, we consider we have
t enhancement in human-computer
uting bandwidth and development

tructive criticism, collaborative re-
, Arthur James, Peter Litwinowicz,
in the Advanced Technology Group

trademarks of Apple Computer, Inc.

he [*t*] sound in 'train' is different from the
lophones do not change the meaning of a
erent phonetic contexts.
peech segments to build syllables, words,
n diphones, to demi-syllables, to whole

ıone). See **Length**.
strong regional accent, and is typified by

or sentence. E.g. the statement "You are
nation patterns, or tunes.
d in milliseconds (ms). For example, the
han the vowel in 'cat', when both words

i.e., concrete realizations of **phonemes**.

Phoneme: any sound that can change the meaning of a word. A phoneme is an abstract unit that encompasses all the pronunciations of similar context-dependent variants. A phonemic representation is commonly used to encode the transition from written letters to an intermediate level of representation that is then converted to the appropriate sound segments (**allophones**).

Pitch: the perceived property of a sound or sentence by which a listener can place it on a scale from high to low. Pitch is the perceptual correlate of the fundamental frequency, i.e., the rate of vibration of the vocal folds. Pitch movements are effected by falling, rising, and level contours. Exaggerated speech, for example, would contain many high falling pitch contours, and bored speech would contain many level and low-falling contours.

Pitch range: the variation around the average pitch, the area within which a speaker moves while speaking in intonational contours. Pitch range has a median, an upper, and a lower part.

Prosody: a collective term used for the variations that can occur in the **suprasegmental** elements of speech, together with the variations in the **rate** of speaking.

Rate: the speed at which speech is uttered, usually described on a scale from fast to slow, and measured in words per minute (wpm). Allegro speech is fast and legato speech is slow. Speaking rate will contribute to the perception of the **speech style**.

Semitone: a pitch interval halfway between two whole tones. There are 12 semitones in an octave. A semitone scale is non-linear and interval-preserving. The formulae for converting semitones to Hertz and vice versa are given in *Inside Macintosh. Sound* (1994, pp. 4–7).

Speaking fundamental frequency: the average (mean) pitch frequency used by a speaker. May be termed the 'baseline pitch'.

Speech style: the way in which an individual speaks. Individual styles may be clipped, slurred, soft, loud, legato, etc. Speech style will also be affected by the context in which the speech is uttered, e.g., more and less formal styles, and how the speaker feels about what they are saying, e.g., relaxed, angry or bored.

Stop consonant: any sound produced by a total closure in the vocal tract. There are six stop consonants in General American English, that appear initially in the words 'pin, tin, kin, bin, din, gun'.

Suprasegmental: a phonetic effect that is not linked to an individual speech sound such as a vowel or consonant, and which extends over an entire word, phrase or sentence. Rhythm, **duration**, **intonation** and **stress** are all suprasegmental elements of speech.

Vocal cords: the two folds of muscle, located in the larynx, that vibrate to form voiced sounds. When they are not vibrating, they may assume a range of positions, going from closed tightly together and forming a glottal stop, to fully open as in quiet breathing. Voiceless sounds are produced with the vocal cords apart. Other variations **pitch** and in **voice quality** are produced by adjusting the tension and thickness of the vocal cords.

Voice quality: a speaker-dependent characteristic which gives a voice its particular identity and by which speakers are most quickly identified. Such factors as age, sex, regional background, stature, state of health, and the overall speaking situation will affect voice quality; e.g., an older smoker will have a creaky voice quality; speakers from New York City are thought to have more nasalized voice qualities than speakers from other regions; a nervous speaker may have a breathy and tremulous voice quality.

Volume: the overall amplitude or loudness at which speech is produced.

References

1. J. Allen, M.S. Hunnicutt, and D. Klatt, From Text to Speech: The MITalk System, Cambridge University Press: Cambridge, 1987.
2. C. Baber, "Speech output," in Interactive Speech Technology, C. Baber and J.M. Noyes (Eds.), Taylor and Francis: London, 1993, pp. 21–24.
3. B.L. Brown, W.J. Strong, and A.C. Rencher, "Fifty-four voices from two: The effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech," Journal of the Acoustical Society of America, Vol. 55, pp. 313–318, 1974.
4. J.E. Cahn, "Generating expression in synthesized speech," Technical Report, M.I.T. Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1990.
5. R. Carlson, B. Granström, and I. Karlsson, "Experiments with voice modelling in speech synthesis," Speech Communication, Vol. 10, pp. 481–489, 1991.
6. R. Collier, "Multi-language intonation synthesis," Journal of Phonetics, Vol. 19, pp. 61–74, 1991.

7. C.K. Cowley and D.M. Jones, "Assessing the quality of synthetic speech," in Interactive Speech Technology, C. Baber and J.M. Noyes (Eds.), Taylor and Francis: London, 1993, pp. 149–155.
8. D. Crystal, The English Tone of Voice, Edward Arnold: London, 1975.
9. Digital Equipment Corporation, DECtalk DTC03 Text-to-Speech System Owner's Manual, Maynard, MA, 1985.
10. J.H. Eggen, "On the Quality of Synthetic Speech, Evaluation and Improvements," Doctoral Thesis, University of Eindhoven, 1992.
11. R.W. Frick, "The prosodic expression of anger: Differentiating threat and frustration," Aggressive Behavior, Vol. 12, pp. 121–128, 1986.
12. C.G. Henton, "Fact and fiction in the use of female and male pitch," Language and Communication, Vol. 9, pp. 299–311, 1989.
13. C. Henton, "The abnormality of male speech," in New Departures in Linguistics, G. Wolf (Ed.), Garland Press: New York, 1992a, pp. 27–58.
14. C. Henton, "Sex and speech synthesis: Techniques, successes, and challenges," in Proceedings of the Fourth Australian International Conference on Speech Science and Technology (SST-92), Brisbane, 1992b, pp. 738–743.
15. C. Henton, "Speech synthesis: Telling it like it is," Australasian Wheels for the Mind, Vol. 3, pp. 40–45, 1993.
16. C. Henton, "Beyond visemes: Using disemes in synthetic speech with facial animation," Journal of the Acoustical Society of America, Vol. 95, p. 3010, 1994.
17. C. Henton, "Pitch dynamism in female and male speech," Language and Communication, Vol. 15, pp. 43–61, 1995.
18. C. Henton and P. Litwinowicz, "Saying it with feeling: Techniques for synthesizing visible, emotional speech," in Proceedings, 2nd. ESCA/IEEE Workshop on Speech Synthesis, 1994, pp. 73–76.
19. *Inside Macintosh. Sound* (1994), Apple Computer, Inc., Cupertino, CA.
20. A. James and J.C. Spohrer, "Simulation-based learning systems: Prototypes and experiences," in Proceedings, ACM/SIGCHI Human Factors in Computing Systems, Monterey, CA, May 3–7, 1992, pp. 523–524.
21. D.H. Klatt, "Review of text-to-speech conversion for English," Journal of the Acoustical Society of America, Vol. 82, pp. 737–793, 1987.
22. D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," Journal of the Acoustical Society of America, Vol. 87, pp. 820–855, 1990.
23. J. Laver, The Phonetic Description of Voice Quality, Cambridge University Press: Cambridge, 1980.
24. P. Litwinowicz and L. Williams, "Animating images with drawings," SIGGRAPH '94 Conference Proceedings, 1994, pp. 121–124.
25. D.W. Massaro, "Speech perception by ear and by eye: A paradigm for psychological enquiry," Lawrence Erlbaum Associates: Hillsdale, NJ, 1987.
26. D.W. Massaro, M.M. Cohen, and P.M.T. Smeele, "Cross-linguistic comparisons in the integration of visual and auditory speech," Memory and Cognition, Vol. 23, pp. 113–131, 1995.
27. D.W. Massaro and E.L. Ferguson, "Cognitive style and perception: The relationship between category width and speech perception, categorization, and discrimination," American Journal of Psychology, Vol. 106, pp. 25–49, 1993.
28. I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," Journal of the Acoustical Society of America, Vol. 93, pp. 1097–1108, 1993.
29. A. Ortony and T.J. Turner, "What's basic about basic emotions?," Psychological Review, Vol. 97, pp. 315–331, 1990.
30. D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley: Reading, Mass., 1990.
31. E. Patterson, P. Litwinowicz, and N. Greene, "Facial animation by spatial mapping," Computer Animation 1991, Springer Verlag: New York, 1991, pp. 31–44.
32. K.R. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," Review of Personality and Social Psychology, Vol. 5, pp. 37–63, 1984.
33. J.C. Spohrer, A. James, C.A. Abbott, G.J. Czora, J. Laffey, and M.L. Miller, "A role-playing simulator for needs analysis consultations," in Proceedings of the World Congress on Expert Systems, Pergamon Press: Orlando, FL, 1991.
34. K.N. Stevens and C.A. Bickley, "Constraints
Journal of Phonetics, Vol. 19, pp. 161–174,
35. M. Tatham, "Voice output for human-mach
J.M. Noyes (Eds.), Taylor and Francis: Lon
36. R.A.M.G. van Bezooijen, Characteristics
Dordrecht, 1984.
37. T. Vitale, "Issues in speech technology fo
Society, Vol. 12, pp. 13–34, 1992.
38. E.J. Yannakoudakis and P.J. Hutton, Speecl
1987.

Caroline Henton received her doctorate in Phon
of Oxford, the University of Sheffield, Universi
Oxford and UCLA, and gave invited courses a
Dutch LOT Winterschool. She left academia in
she created the high quality synthetic speech on
head with synthetic speech. 1994–1995 she cor
and text-field localization rules in 7 languages.
Sound Corp., Lexicon Naming, and for Apple.

She joined Voice Processing Corporation in
ordinating foreign language ASR projects, one
interface for an 'automated receptionist' CTI ap
product management aspects of the SUI are al
publications, focusing on topics in acoustic ph
between male and female speech.

Brad Edelman received a BS in Computer S
in computer graphics and application framewo
Associates and Apple Computer; he has worked
Laboratory (UBILAB). He is currently an empl
products, including Adobe PageMill, a WYSIW

tic speech," in Interactive Speech Technology,
1993, pp. 149-155.
)n, 1975.
ech System Owner's Manual, Maynard, MA,

d Improvements," Doctoral Thesis, University

; threat and frustration," Aggressive Behavior,

)itch," Language and Communication, Vol. 9,

tures in Linguistics, G. Wolf (Ed.), Garland

and challenges," in Proceedings of the Fourth
Technology (SST-92), Brisbane, 1992b, pp.

ian Wheels for the Mind, Vol. 3, pp. 40–45,

)eech with facial animation," Journal of the

iage and Communication, Vol. 15, pp. 43–61,

:s for synthesizing visible, emotional speech,"
:is, 1994, pp. 73–76.
no, CA.
Prototypes and experiences," in Proceedings,
y, CA, May 3-7, 1992, pp. 523–524.
ournal of the Acoustical Society of America,

)f voice quality variations among female and
. 87, pp. 820–855, 1990.
University Press: Cambridge, 1980.
s," SIGGRAPH '94 Conference Proceedings,

digm for psychological enquiry," Lawrence

:tic comparisons in the integration of visual
131, 1995.
)n: The relationship between category width
erican Journal of Psychology, Vol. 106, pp.

1 synthetic speech: A review of the literature
\merica, Vol. 93, pp. 1097–1108, 1993.
Psychological Review, Vol. 97, pp. 315–331,

ie, Addison-Wesley: Reading, Mass., 1990.
by spatial mapping," Computer Animation

l and some cross-cultural data," Review of

M.L. Miller, "A role-playing simulator for
gress on Expert Systems, Pergamon Press:

34. K.N. Stevens and C.A. Bickley, "Constraints among parameters simplify control of Klatt formant synthesizer," Journal of Phonetics, Vol. 19, pp. 161–174, 1991.
35. M. Tatham, "Voice output for human-machine interaction," in Interactive Speech Technology, C. Baber and J.M. Noyes (Eds.), Taylor and Francis: London, 1993, pp. 25–35.
36. R.A.M.G. van Bezooijen, Characteristics and Recognizability of Vocal Expressions of Emotion, Foris: Dordrecht, 1984.
37. T. Vitale, "Issues in speech technology for persons with disabilities," Journal of the American Voice I/O Society, Vol. 12, pp. 13–34, 1992.
38. E.J. Yannakoudakis and P.J. Hutton, Speech Synthesis and Recognition Systems, Halsted Press: New York, 1987.

Caroline Henton received her doctorate in Phonetics from the University of Oxford; she has taught at the University of Oxford, the University of Sheffield, University of California Davis and UCSB. She held research positions at Oxford and UCLA, and gave invited courses at the Linguistic Society of America Linguistic Institute and the Dutch LOT Winterschool. She left academia in 1990 to join Apple Computer as a Senior Research Scientist; there she created the high quality synthetic speech on Apple computers as well as collaborated on an animated talking head with synthetic speech. 1994–1995 she consulted for Sun Microsystems, developing phonetic specifications and text-field localization rules in 7 languages. She has also consulted for Interval Research, Claris Corp., Digital Sound Corp., Lexicon Naming, and for Apple.

She joined Voice Processing Corporation in 1995 as Director of Language Development. Together with co-ordinating foreign language ASR projects, one of her responsibilities is the design and testing of a speech user interface for an 'automated receptionist' CTI application. The linguistic, discourse, localization, user testing and product management aspects of the SUI are all issues with which she is concerned. She has written over 40 publications, focusing on topics in acoustic phonetics, speech synthesis, SUI design, and phonetic differences between male and female speech.

Brad Edelman received a BS in Computer Science and engineering from MIT in 1993. His background is in computer graphics and application frameworks. He has held internships at the MIT Media Lab, R/Greenberg Associates and Apple Computer; he has worked full-time at Taligent and the Union Bank Switzerland's Information Laboratory (UBILAB). He is currently an employee of Adobe Systems where he is developing internet publishing products, including Adobe PageMill, a WYSIWIG HTML editor.

For information about current subscription rates and prices for back volumes for *Multimedia Tools and Applications, ISSN 1380-7501* please contact one of the customer service departments of Kluwer Academic Publishers or return the form overleaf to:

Kluwer Academic Publishers, Customer Service, P.O. Box 322, 3300 AH Dordrecht, the Netherlands, Telephone (+31) 78 524 400, Fax (+31) 78 183 273, Email: services@wkap.nl

or

Kluwer Academic Publishers, Customer Service, P.O. Box 358, Accord Station, Hingham MA 02018-0358, USA, Telephone (1) 617 871 6600, Fax (1) 617 871 6528, Email: kluwer@world.std.com

Call for papers

Authors wishing to submit papers related to any of the themes or topics covered by *Multimedia Tools and Applications* are cordially invited to prepare their manuscript following the 'Instructions for Authors'. Please request these instructions using the card below.

Author response card

Multimedia Tools and Applications

I intend to submit an article on the following topic:

_____

_____

Please send me detailed 'Instructions for Authors'.

NAME : _____

INSTITUTE : _____

DEPARTMENT : _____

ADDRESS : _____

_____

_____

Telephone : _____

Telefax : _____

Email : _____

Library Recommendation Form

Route via Interdepartmental Mail

To the Serials Librarian at: _____

From: _____ Dept./Faculty of: _____

Dear Librarian,

I would like to recommend our library to carry a subscription to **Multimedia Tools and Applications**, ISSN 1380-7501 published by Kluwer Academic Publishers.

Signed: _____ Date: _____

Generating and Ma
Speech on a Person

CAROLINE HENTON
Voice Processing Corporation, 1 Main St

BRADLEY EDELMAN
Internet Products Group, Adobe Systems

Abstract. Against a background of inc
proposed for users of the simulator and
vocal emotion in synthetic speech using
synthesizer. The second enhancement al
output by the text-to-speech system. Voc
by the user. Applications such as traini
the addition of emotions. A graphical e
authoring environment of these applicati

Keywords: emotions in synthetic spee

1. Introduction

The central question we attem
'talking head' appear more hum
an authoring environment for pro
that can be manipulated using a

At the outset we give the br
tool. Next, we review the literat
that have any ability to simulate
ing a limited number of prosodi
produced with a diphone-conca
thesizer is the one included in th
first released on the Apple Maci

We give a detailed account of
and allows for their direct contro
synthetic speech, the approach
high level of abstraction. A use
editor will sound because of the

For reasons of logic and cl
first is concerned with the spe
interface. This order of explana

Request for information about current subscription rates and prices for back volumes of

Multimedia Tools and Applications, ISSN 1380-7501

Please fill in and return to:

Kluwer Academic Publishers, Customer Service, P.O. Box 322, 3300 AH Dordrecht, the Netherlands

Kluwer Academic Publishers, Customer Service, P.O. Box 358, Accord Station, Hingham MA 02018-0358, USA

☐ Please send information about current program and prices
☐ Please send a free sample copy

NAME :
INSTITUTE :
DEPARTMENT :
ADDRESS :
Telephone :
Telefax :
Email :

REF. OPC

STAMP

Multimedia Tools and Applications
Kluwer Academic Publishers,
101 Philip Drive
Assinippi Park
Norwell, MA 02061
U.S.A.

TO : The Library
FROM: _____

VIA INTERDEPARTMENTAL MAIL